MQB7046: MODELLING PUBLIC HEALTH DATA

Semester 2, Session 2023/2024

CONTINUOUS ASSESSMENT 4

Please read the instructions carefully. Failure to comply with any of these instructions may result in penalty of marks or grading of this assessment.

1. This assignment represents **30%** of the evaluation for this course.
2. Answer ALL questions in English.
3. This assignment is "open book," which means you are permitted to use any materials handed out in class, your own notes from the course and textbooks.
4. You are allowed to submit this assignment in a group of maximum 3 persons.
5. You need to submit a report in MSWord that contain introduction, methods, results, analysis, discussion and conclusion.
6. All analyses should be conducted using the Python software. All codes and outputs need to be interpreted.
7. Submission instructions:
    - For students found plagiarising or commit any forms of academic misconducts, they will be penalised and given a zero (fail).
    - Upload the following; 1) your report in MSWord; 2) your jupyter notebook; 3) Any dataset or appendix deemed necessary. Submit in a folder and upload in the SPeCTRUM
    - All documents must be named according to the Matrix_number
    - Submit your files via SPeCTRUM before the due date on **10/6/2024 (Monday) 12.00pm.** Failure to submit your assignment on the stated due date and time will be given a zero.

Question 1.

This dataset contains records of 553 patients, encompassing a mixture of individuals diagnosed with and without liver disease. The prediction task is to determine whether a patient suffers from liver disease based on information about several biochemical markers, as well as clinical and sociodemographic profiles. The dataset comprises 12 features and 1 target variable. The details of the dataset (assignment.csv) are as follows:

Variable Name	**Description**
Age	Age
gender	Gender
ethnic	Patient's ethnicity: 1-Malay, 2-Chinese, 3-Indian
TB	Total Bilirubin
DB	Direct Bilirubin
TP	Total Proteins
ALB	Albumin
SGPT	Alanine Aminotransferase
SGOT	Aspartate Aminotransferase
Alkphos	Alkaline Phosphatase
AGR	Albumin and Globulin Ratio
BMI	Body mass index: 1- Normal, 2- Overweight
Disease	Liver disease: 1 – No disease, 2 - Disease

The assignment should follow this overall outline:

1. Introduction - describe the background, motivation and importance of the data in light of related literature. Provide any descriptive statistics of the contained variables. [10 marks]
2. Methodology
    - Data preprocessing - describe details of preprocessing of the data, including data cleaning, data imputation, etc of the dataset.
    - Algorithm design and implementation - Select at least two ML models, provide description of both algorithms including their rationale or model structure, generate training and testing data with appropriate format that suit the ML model, and model evaluation based on the outcome.
    - Demonstrate your solution with an attached iPython notebook. Ensure reproducibility and transparency. [50 marks]
3. Results – Present results and any evaluation criteria such as a confusion matrix, precision, recall, RMSE, MARD, F1, AUC and a ROC-plot, etc. Provide a sensitivity analysis for both algorithms with different parameters and give a textual description of the results. [20 marks]
4. Discussion and Conclusion

- Compare and discuss your findings (results of two algorithms) with existing literature.
- Discuss how you would improve your methodology, current limitations and future work. [20 marks]

5. Reference (any relevant references)
6. Appendix: Attach a reproducible iPython Jupiter notebook.